DOCUMENT RESUME

ED 270 503 TM 860 393

AUTHOR Wilson, Ann Jarvella
TITLE Historical Issues of Validity and Validation: The National Teacher Examinations.
PUB DATE Apr 86
NOTE 32p.; For the history of the National Teachers Examinations Program, see ED 026 049.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; Communication Skills; Competency Based Teacher Education; Court Litigation; Educational Background; *Educational History; Minimum Competency Testing; National Competency Tests; *Occupational Tests; *Teacher Certification; Teacher Education; Teacher Education Curriculum; *Teacher Evaluation; Teachers; Test Format; Testing Problems; *Test Use; *Test Validity
IDENTIFIERS *National Teacher Examinations

ABSTRACT

The purpose of this paper, which is drawn from a larger analytic history of the National Teacher Evaluation (NTE) program, is to investigate issues of validity within the context of the program's 50-year history. Three major findings emerge from historical considerations relating to: (1) the continuity of test content and justification over the 50-year period of the program's existence; (2) the primacy of reliance upon logical or content validity; and (3) the paradoxical relationship of the tests to teacher education curricula. First, since their inception the examinations have measured three categories of teacher knowledge--basic intellectual and communicative skills, general cultural and contemporary background, and pedagogical and professional information. When changes did occur, they were undertaken for either financial reasons or as responses to specific criticisms. Second, there has been a strong tendency to justify the exams in terms of their logical or practical validity. Finally, despite the persistent assumption that the tests are needed because graduates of many teacher education programs are inadequately prepared, the source of test content and validity has been and continues to be focused primarily upon the perceived curricula of those programs. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270503

HISTORICAL ISSUES OF VALIDITY AND VALIDATION:

THE NATIONAL TEACHER EXAMINATIONS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. J. Wilson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Ann Jarvella Wilson

Department of Education
Carroll College
100 North East Avenue
Waukesha, Wisconsin 53186

Paper presented at the Annual Meeting
of the American Educational Research Association
San Francisco, April 1986

TM 860 393

DRAFT COPY:
Do not quote or copy without permission.
Comments are welcome.

Historical Issues of Validity and Validation: The National Teacher Examinations

Introduction

In the past several years, concerns about school quality and teacher competence have focused public and professional attention on tests for teachers, most often on the National Teacher Examinations. This battery of standardized tests is currently used for teacher assessment and/or certification in some twenty-five states. It has also served as the model for the California Basic Educational Skills Tests and the Pre-Professional Skills Tests--exams used in California, Texas, and elsewhere for admission into teacher education programs. These tests were originally developed in the 1930's and are currently prepared by the Educational Testing Service. Reviewers in the Buros's publications have periodically criticized the tests' lack of empirically documented validity, but for the most part, neither the exams' content nor their validation have received much critical attention. Only in the past decade have legal challenges to the tests' use and documentation of their negative impact on minority teachers focused more than passing attention to issues of validity and validation.

The purpose of this paper, which is drawn from a larger analytic history of the NTE program,[1] is to investigate issues of validity within the context of the program's fifty year history. Drawing upon the literature of the sociology of school knowledge, sources of test validity [the relationships between a test and what it purports or is designed to measure] and methods for test validation [the procedures for documenting those relationships] are explored by relating the content and construction of successive versions of the examinations[2] to (1) justifications for test use and assumptions about validity made by program administrators, (2) validation procedures and techniques recommended by test officials, and (3) major validity studies conducted by project and/or independent researchers.

Historical Concepts of Test Validity

Concepts of test validity have been evolving since the early part of the twentieth century. "The earliest writings on the subject recognized two types of validity, logical and experimental."[3] The later involved "comparison of the results secured on [a] test . . . with those obtained from other measures of the same thing," and the former was based upon "the careful inspection and analysis of the test itself."[4] Logical or practical validity was that "built into the test" by careful

planning and construction, sometimes with the use of explicit and comprehensive descriptive rationales.[5] By the late 1930's, professional attention focused upon "direct measurement as a means of attaining validity,"[6] and experimentally determined validity was emphasized. Now often seen as "the step-child of testing,"[7] contemporary logical or content-related validity is concerned with the "degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content."[8] Although practical validation continues to be promoted by some test theorists,[9] most measurement experts have favored the collection of empirical data and the correlation of test scores with criterion measures.[10]

The initial technical standards prepared by the American Psychological Association (APA) in the early 1950's[11] recognized four distinct types of validity--(1) content validity involving "the sampling of a specified universe of content;" (2) concurrent validity involving "the relation of test scores to an accepted contemporary criterion of performance;" (3) predictive validity involving "the relation of test scores to measures [taken] at some later time;" and (4) construct validity involving "more indirect validation procedures . . . "[12] In the 1966 revision of the standards,[13] the predictive and concurrent categories were merged and treated as alternative forms of criterion-related validity.

In recent years, construct validity with its concern for "understanding the underlying dimensions or attributes being measured,"[14] has come to be seen as a unifying concept which subsumes all other types of validity.[15] The most recent APA standards still differentiate between content-related, criterion-related, and construct-related "evidences" of validity, but they state that gathering construct-related evidence "begins with test development and continues until the pattern of empirical relations between test scores and other variables clearly indicates the meaning of test score."[16] Thus in many ways, "all validation is one, and in a sense all is construct validation."[17]

Antecedents of the National Teacher Examinations

Although exams for teachers have been used in the United States since colonial times, reliance upon them diminished with the development and expansion of teacher training programs. In the late 1920's and early 1930's,[18] nationwide emphases on school efficiency and accountability fueled by thriving intelligence and achievement testing and accompanied by a concurrent teacher surplus gave teacher testing new momentum. Research bureaus affiliated with urban school districts or with colleges and universities often constructed local tests for teaching candidates

along with those for school children,[19] and several batteries of tests for teachers were sold nationwide.[20] Within this context, Pennsylvania launched a state-wide educational study which, though not originally designed to test teaching candidates, led directly to the National Teacher Examinations.

Beginning in 1925, the Pennsylvania study was funded by the Carnegie Corporation to evaluate the quality of and relationships between the states's secondary and higher education systems.[21] In charge of the project were William Learned of the Carnegie Foundation staff and Ben Wood, a national authority on objective testing and the director of Collegiate Educational Research at Columbia University. In 1928, graduating seniors in Pennsylvania's high schools were given a massive battery of commercial intelligence and achievement tests. That same year, special twelve-hour exams developed by Wood were administered on a trial basis to the state's college seniors. After revision, these exams were given twice to those 1928 high school graduates who went on to college in Pennsylvania--in 1930 and again in 1932. Selected groups of high school seniors were also tested. Containing matching, true-false, and multiple-choice items, the exams were designed to assess intelligence, English, mathematics, and general culture.

The testers assumed that these exams measured "significant aspects of liberal arts education" and that their validity was demonstrated both by the "scope, distribution, and character of the questions" and by "feasible external checks."[22] Scores showed gains over the two year period for most students at most institutions and correlated reasonably well with college grades.[23]

The major finding of the Pennsylvania study was great variability in tested knowledge, variability which was exhibited among individuals and among institutions as well as within departments in the same institutions. Neither college attendance, nor class placement, nor school grades necessarily corresponded to knowledge displayed on the tests. The findings and interpretations of the Pennsylvania study led eventually to the creation of the American Council on Education's Cooperative Testing Service and to the development of secondary school and college guidance and testing programs.

Though not an original focus of the research, the results of the Pennsylvania study became widely used to decry the academic quality of teachers and teacher candidates.[24] Prospective teachers had tested particularly poorly. Their average scores were the among the lowest of the the entire sample. Learned and Wood devoted one chapter of their final report to an analysis of the teachers' achievement and concluded that "teaching attracts college students who vary widely in the

fundamental quality of their abilities and who fall below a knowledge minimum in a large proportion of cases."[25] Although the authors stated that the eventual solution would require better programs and higher standards in the preparatory institutions, they put considerable emphasis upon the continued use of exams. They recommended that, prior to employment, school authorities test prospective candidates in order to "secure the best possible teachers for the money they have to pay."[26]

The Original National Teacher Examinations

The Cooperative Test Service of the American Council on Education began operations in 1930--partially to prepare tests for the Pennsylvania study. Funded by a ten year grant from John D. Rockefeller and directed by Ben Wood, the service was expected to develop multiple comparable forms of academic high school and college tests.[27] Beginning in 1932, special editions of its exams were prepared and sold for use in teacher selection. By the late 1930's, the Service provided new versions yearly to some fifteen or twenty cities including Providence, Philadelphia, Pittsburgh, and Cleveland.[28] When the subsidizing grant expired, the superintendents sought additional foundational support. Again, as in Pennsylvania, the Carnegie Corporation provided the funds.

In 1939, the American Council of Education established the National Teacher Examinations program to assist school administrators with teacher selection. A committee composed primarily of urban school superintendents whose systems had used the earlier tests was selected by the Council and charged with responsibility for the program.[29] Under the supervision of Ben Wood as project director, the tasks of constructing, administering, and correcting the exams were assigned to the Cooperative Test Service.

The initial tests were prepared following procedures originally used by Wood in Pennsylvania. Staff editors developed preliminary test outlines and tentative item specifications. General suggestions were gathered from the advisory committee and other administrators and supplemented with data gleaned in analyses of "courses of study, textbooks, journal articles, and reports of professional organizations."[30] Outlines were sent to teacher education and school system personnel for review and criticism. Tentative items were tried out in several teacher training institutions.

The original "common" exams,[31] first administered in 1940, assessed that knowledge selected by the administrators as representative of what "all good teachers should know"[32]--basic intellectual and communicative skills, cultural and contemporary background, and professional information. Modeled closely after

those developed for use in Pennsylvania, the exams were multiple choice in nature and emphasized aspects of general and contemporary culture, rather than pedagogy or professional knowledge. The 1940 exam took eight hours and was composed of eleven separate tests. The titles and contributing portions to the "common examination total score" were as follows:

Intellectual and Communicative Skills	30 percent
1. Reasoning	10 %
2. English Comprehension	10%
3. English Expression	10%
Cultural and Contemporary Background	40 percent
4. Contemporary Affairs	10%
Test of General Culture:	
5. Current Social Problems	5%
6. History and Social Studies	5%
7. Literature	5%
8. Fine Arts	5%
9. Science	5%
10. Mathematics	5%
11. Professional Information	30 percent
Education and Social Policy	7.5%
Child Development and Educational Psychology	7.5%
Guidance and Individual and Group Analysis	7.5%
Elementary or Secondary School Methods	7.5%

Announcements for the project emphasized varied standards in teacher preparation institutions and the complex nature of good teaching. The exams, it was stressed, would help select the best candidates from a surplus which varied widely in ability and training. It was also suggested that "the opportunity to 'register' talents on a national scale " would be advantageous to candidates and institutions preparing teachers.[33]

Advertised nationally, test promotion was most successful in the urban areas of the New England and Middle Atlantic states where the practice of examining teaching candidates was already established and where a substantial surplus of teacher candidates existed. In a presentation to other urban administrators, Alexander Stoddard, Philadelphia's superintendent of schools and the chairman of the testing program's advisory committee, stressed the efficiency with which the

exams had selected candidates for only "a few scattered appointments in the past three years" from a waiting list of over 3000 "qualified" applicants.[34] Program announcements emphasized that participation in the national program would save the time and expense of constructing, administering, and scoring local tests.[35] The exams were promoted as the most accurate and economical device known for measuring "essential elements of teaching ability."[36]

Early Considerations of Validity

From the beginning, program officials stressed that the exams did not measure the totality of teaching ability and therefore should not be judged by their correlation with "available criteria" of teaching ability. In an early talk to teacher educators, Ben Wood argued against "the naive error" of judging validity in terms of correlation with measures of teaching success. He likened the tests to physicians' thermometers and stethoscopes--valid instruments but not sufficient for a "complete diagnosis."[37]

Early critics of the exams--many of whom were teacher educators[38]--did, however, raise questions about their validity. One saw the test makers' disclaimers as admissions that "the really important things in teacher selection" were not being measured.[39] Another suggested that, rather than beginning with a definition of "good teaching," the test makers had asked: "What test items of the kind suggested by school superintendents can we devise which will yield answers that are statistically reliable?"[40] Not enough had been done, the critics maintained, to ascertain if persons who could score well on the exams were those also recognized as good teachers.

Test personnel continued to argue that the value of the exams could not be judged by correlating them with "that composite we think of as teaching success."[41] Since teaching ability was a complex combination of numerous interacting factors, it was not "reasonable to expect any one of the essential factors to correlate highly with the total complex."[42] In one much quoted article, Wood suggested that the tests should be judged, instead, by how accurately they measured those parts of teaching they were "designed to measure, namely, intelligence (linguistic and quantitative), general and specific cultures of the types judged desirable by the teacher-selecting authorities, and professional information."[43]

Test personnel stressed that the tests were "constructed by subject matter experts and test technicians so as to insure maximum validity and reliability."[44] In 1940, John Flanagan, associate director of the Cooperative Test Service, carried

out a preliminary empirical study[45] which foreshadowed his later theoretical work on comprehensive test rationales.[46] Flanagan argued that an important type of validity is related to the way a test is constructed. "A test is valid," he stated, "when, according to experts, the sampling of content and mental processes in the test is similar to that indicated in the outline and specifications for the test."[47] This reliance on what was later called content validity--on careful construction and representative content--continued to be stressed in program materials for test users.

Somewhat paradoxically, Flanagan also compared test scores to several commonly "available" measures of teaching ability--supervisors' and students' ratings. Using the test scores of experienced teachers who took the first exams in 1940, he identified twenty-two school systems with employees whose scores differed by at least 100 points. School superintendents were asked to secure both supervisory and pupil ratings for the forty-nine teachers selected. The correlation of supervisors' "overall judgment of the teachers' general effectiveness and desirability" was .51.[48] Correlations with other supervisory ratings were reported as "around .50." Pupil data were not reported in terms of correlation coefficients but suggested a relationship between test scores and student perceptions of teacher characteristics.

Over the next few years, other investigators attempted to assess validity experimentally. Some compared test scores to supervisors' or principals' ratings.[49] Since these "measures" were of such varied reliability, it is not surprising that much of this work was criticized later by more psychometrically sophisticated testing proponents.[50] Later in the decade, investigations were broadened to include comparisons with concurrent and predictive measures of achievement in college or graduate school.[51] Although not directly discouraging this kind of research, both early and later program personnel tended to attribute the low to moderate correlations yielded by these studies to their theoretical or technical inadequacies.

Financial Distress and Temporary Solutions: The NTE Program in the 1940's

With the second world war came a severe reduction in the number of applicants for teaching positions. The oversupply of teachers dissipated and so did the market for the examinations. Under the leadership of David Ryans,[52] director of the Cooperative Test Service in the middle 1940's, the NTE program managed to survive by incorporating a number of cost-cutting procedures.

The major response to the adverse financial situation was the abandonment of almost all new test construction and the reuse of those exams already prepared. For the first three years of the program's operation, original and comparable forms of the entire exam had been constructed annually by the Cooperative Test Service. However, during 1943, 1944, and 1945 only the Contemporary Affairs section of the common battery was newly prepared each year.[53] Refurbished versions of the exam were provided by combining sections of the earlier three tests.

In 1944, Ryans convinced the national advisory committee that this procedure could not continue indefinitely. The committee approved the reorganization of the general culture component and shortened the common exam enough to be administered in a single day, a move which resulted in lower administrative costs.[54] To supplement his meager staff, Ryans found outside specialists--many of whom were affiliated with the testing bureaus of midwestern colleges and universities--willing to help prepare and review the exams.[55] In an attempt to mollify teacher educators, the weighting of the professional section of the NTE was modified. In addition, each of the professional tests began to be reported separately on the score profile "for guidance purposes."[56] Thus, beginning in 1946, the titles and contributing portions to the "common examination total score" were as follows:

Intellectual and Communicative Skills	30 percent
1. Reasoning	10 %
2. English Comprehension	10%
3. English Expression	10%
Cultural Background	30 percent
4. History, Literature, and Fine Arts	10%
5. Science and Mathematics	10%
6. Contemporary Affairs	10%
Professional Information	40 percent
7. Education and Social Policy	10%
8. Child Development and Educational Psychology	10%
9. Guidance and Individual and Group Analysis	10%
10. General Principles and Methods of Teaching	10%

In order to save money, the number of items and total testing time were further reduced each year until 1950, the final year that the project was affiliated with the American Council on Education. Even with these modifications, the exams used at the end of the decade were very similar to those originated in 1940.

During and following the war, additional efforts were made to secure support and broader use within the teacher education community. Both the composition and the leadership of the national advisory committee were changed to include more representation by teacher training personnel. Reduced student fees were offered "to acquaint colleges and students with the program,"[57] and promotional materials aimed at teacher education personnel were prepared.[58] In spite of these moves, however, test use by students remained very low, and it became clear that other sources of income would be needed. Supplementary grants from the Carnegie Corporation in 1940 and 1941 helped offset the war's immediate effects, but no further foundation monies were provided.[59]

Beginning in 1944, additional revenue was secured from test sales to the State of South Carolina for use in a new teacher certification program. Over the next few years, these test administrations provided the major source of NTE funding.[60] South Carolina's new system relied on NTE scores to determine "grade" of certification (and thus state salary reimbursement) for both experienced and new teachers and replaced a dual system based upon race similar to one which had been outlawed by the U.S. Supreme Court in 1940. A "validation" study conducted by teacher educators at the University of South Carolina with the assistance of Ben Wood as consultant[61] compared selected groups of white teachers and teacher candidates. It concluded that "successful teachers in South Carolina are likely to make higher scores [on the National Teacher Examinations] than prospective teachers who are seniors in the colleges of the State."[62] Subsequent state-wide administrations revealed that white teachers tended to outscore blacks and eventually the system was challenged in the courts. For years, however, NTE scores maintained a salary differential previously based explicitly on race.[63] Although alluded to in one program publication,[64] South Carolina's use of the tests for salary purposes was rarely described in program materials.

For the rest of the decade, NTE informational and promotional materials were prepared by David Ryans. He also wrote most of what was published about test validity, drawing on his and others' earlier work, and usually reiterating familiar arguments. His 1949 article for school administrators[65] was drawn from "The National Teacher Examinations: Notes on the Question of Their Validity," an informational sheet he had prepared and provided to test users in 1946. It reported on "two preliminary statistical studies." The first of these was the Flanagan study, the other a comparison in one unnamed college of prospective teachers' scores with faculty ratings of their "probable success." Never published except as a brief item in Ryans's newsletter for potential exam users,[66] this research apparently was an attempt to provide validity data to justify test use in colleges and universities.

Like his predecessors, Ryans argued that high correlations between the NTE and "the usual criteria of teaching success" were unlikely because "no adequate criteria of teaching success" yet existed and because the exams measured "only one phase of teaching ability." They did, he believed, "provide reliable estimates of the candidates' intellectual and cultural backgrounds."[67] No mention was made of South Carolina's study.

Despite underfunded and inadequate test development for much of this period, Ryans continued to emphasize the tests' content validity and indicated that the major source of their validity lay in the way in which they were prepared. He commended the tests' "constant" revisions and their relationship to "materials that are believed to be important for teachers to know" and concluded that "from the standpoint of their representativeness of types of materials and objectives they are prepared to measure, there is little question of the validity of the Teacher Examinations."[68]

Transitions and Recovery: The NTE Program in the 1950's

Late in 1947, in order to deal "with testing and measurement in a coordinated manner and [to eliminate] duplication of effort,"[69] the American Council on Education merged its testing programs with those of the College Entrance Examination Board and the Graduate Record Office to form a new organization--the Educational Testing Service. Between 1948 and 1951, project administration, test preparation, and eventually sponsorship of the National Teacher Examinations program were transferred to the new agency.[70]

Guided by the overall leadership of Henry Chauncey, president of the Educational Testing Service, and by the specific project direction of Arthur Benson, strenuous efforts were undertaken to economize, to make the program self-supporting and more efficient. Administrative procedures were simplified and the exams "streamlined."[71]

The version of the National Teacher Examinations administered by ETS in 1951 was the shortest and quickest test in the history of the program. The common examination, which in 1940 had included 1217 items to be answered in eight hours of working time, was reduced to three hundred items and a working time of just over three hours. In content and structure, however, the test was remarkably similar to those administered earlier. In fact, many of the items on the cultural and professional sections were taken directly from earlier tests.[72] The test of reading ability was eliminated. Contemporary content from a previously separate subtest

was incorporated into the other general culture sections. A new "weighted common examination total" or "WCET" score was created to be comparable to the earlier total score. Titles and contributing portions to the "WCET" became as follows:

Intellectual and Communicative Skills	20 percent
1. Reasoning	10 %
2. English Expression	10%
Cultural Background	40 percent
3. History, Literature, and Fine Arts	20%
4. Science and Mathematics	20%
5. Professional Information	40 percent
Education as a Social Institution	10%
Child Development and Educational Psychology	10%
Guidance and Measurement	10%
General Principles and Methods of Teaching	10%

Although the basic examinations changed little during the next decade, the program diversified with the development of specialized state- and institution-wide testing services, supplementary tests for administrators and others, and new subject-matter exams. Except for their shortening, however, the scope and the emphases of the common battery during the 1950's resembled those of the earlier exams.

Most research conducted during this period was done by masters and doctoral students and involved the assessment of teaching candidates trained at a particular college. Exam scores were correlated with undergraduate grade point average[73] or with achievement test scores.[74] Correlations were also computed between National Teacher Examinations scores and various assessments of teaching ability.[75] Most of what was published about the exams during the 1950's was prepared by NTE project director, Arthur Benson, who was responsible for both informational and promotional materials. As in the late 1940's, the tests were recommended to teacher educators for "institutional evaluation, counseling and placement activities, and screening . . . for graduate work,"[76] but test use was still justified primarily in terms of widely varied teacher preparation. NTE results were said to be "a useful supplement to academic records since they [provided] school systems with comparable measures for all teacher applicants without regard to the standards of the institutions which prepared them."[77]

The Educational Testing Service conducted no original NTE validation research during this decade but made references to content and concurrent validity in project publications. A statement in the program's first specialized pamphlet for users, the Handbook for School and College Officials, published in 1959, stressed that "a priori evidence as to content validity . . . is inherent in the manner in which the tests are planned and constructed."[78] Potential users were encouraged to "inspect the tests to determine the relevance of the test materials to their [own] purposes." Although no specific references were cited, the handbook stated that "periodic reports of studies which have related NTE scores to such criteria as grade point averages or credit hours of collegiate study have been consistent in supporting the [tests'] concurrent validity."[79]

Predictive validation was presented as problematic. Benson repeatedly criticized "so-called validity studies"[80] which attempted to measure the tests "against on-the-job performance"[81] and argued that "vaguely defined ratings by supervisors or administrators" were no longer acceptable "as adequate criteria of teacher effectiveness."[82] A further statement about "on-the-job" criteria appeared first as a footnote in the 1961 publication for users[83] and then as part of the text in the 1964 version.[84] It read: "The validity of the NTE is more appropriately judged on the basis of proximate criteria than on ultimate success in teaching. Until research establishes universally acceptable criteria of teaching effectiveness, results of validating the NTE against on-the-job performance of teachers are likely to be inconclusive. . . . " In 1967, the statement was modified to blame the lack of predictive criteria on "professional educators:" "At present, professional educators are unable to agree on the meaning of 'teaching effectiveness.' Until educators are able to define and divide this criterion into components which can be validly and reliably measured, this method of substantiating or refuting the validity of the NTE will remain relatively unsuccessful."[85]

Professionalizing the Examinations: The NTE's in the 1960'S

Like teacher education a decade earlier, the National Teacher Examinations became the focus of growing critical attention in the 1960's. The ascent of Sputnik, the poor showing of teachers on the Selective Service Qualifying Tests, concern about the alleged dominance of teacher preparation by "educationists,"[86] and an intense debate over the relationship between general and professional components of teacher education--all affected attitudes toward teacher preparation and toward teacher tests. In 1961, an external review committee, nominated by the National Education Association's National Commission on Teacher Education and Professional Standards, recommended extensive alternations in the organization and the content

of the exams. Although advocating test use "as an aid in teacher selection," the review committee recommended the establishment of "new norms based on a nationwide sampling of all prospective teachers"[87] and discouraged "the use of scores for other purposes, such as certification" until revisions were made. It also called for periodic program review and for involvement of additional persons not affiliated with ETS to help plan, write, and review the tests.

These changes, many of which emphasized "professionalizing" the knowledge assessed on the tests, were implemented for the 1964-65 testings and involved the first major revisions of the exams since the Educational Testing Service took over the project more than a decade before. Eliminated at last was the nonverbal reasoning test which the committee had believed had "no particular relevance" for testing teachers and could not "be considered a test purported to measure academic preparation."[88]

A new publication for teacher examination users, Prospectus for School and College Officials, was prepared to "aid school and college officials . . . in making judgments regarding the appropriateness of the National Teacher Examinations program for their particular measurement needs and to assist them in planning to use the results of these examinations effectively."[89] While noting that the question of "what knowledge is of most worth to prospective teachers?" was considered in exam development, the booklet stressed that the program provided "objective exams of measurable knowledges and abilities which [were] commonly considered basic to effective classroom teaching and which typically [constituted] major elements in current programs of teacher education."[90] This concentration on teacher preparation programs as a source of the knowledge tested was emphasized in another new publication, the Technical Handbook, which appeared in 1965. It announced that "the chief purpose of the NTE is to provide an independent evaluation of the academic preparation of teacher education students."[91]

The new battery was organized as a set of three general education tests and three professional educational tests. The titles and contributions to the new weighted common examination total were as follows:

General Education	61 percent
1. Written English Expression	11%
2. Social Studies, Literature, and the Fine Arts	25%
3. Science and Mathematics	25%

Professional Education	39 percent
4. Societal Foundations of Education	13%
5. Psychological Foundations of Education	13%
6. Teaching Principles and Practices	13%

Two studies conducted by ETS staff in the early 1960's further reinforced the focus on the tests' relationship to teacher education curricula--Barbara Pitcher's study of concurrent test validity[92] and Betty Humphry's survey of professional course offerings.[93] Pitcher, an employee of ETS's Statistical Analysis Division, analyzed test score and grade point data of college seniors who graduated in 1959, 1960, or 1961 from eleven teacher preparatory institutions. Correlations between cumulative grade point averages and weighted common examination total scores ranged from .38 to .74 with a weighted average of .57. She concluded that this represented a reasonably high relation between test scores and college grades. Although published only as an internal statistical report, Pitcher's research was the first NTE validation study undertaken by ETS personnel and for almost two decades was cited to document the exam's concurrent validity.[94]

Head of the Education Section, Test Development Division, and in charge of preparing test specifications for the newly revised exams, Humphry surveyed professional education requirements in some 250 colleges and universities in 1961-62. Finding considerable overlap in course requirements and materials used in institutions approved by the National Council for Accreditation of Teacher Education, she concluded that "there is perhaps more agreement concerning the basic content taught than might seem readily apparent."[95] Though not mentioned as often as Pitcher's work, Humphry's survey was also cited in program publications as partial documentation of the exam's content validity.

Responding to External Pressures: The NTE in the 1970's

Even before the recommendations of the 1961 review committee had been implemented with the restructuring of the 1964-65 common exam, there was growing impetus for further action. Rapid growth in test adoption by state and local school systems in the recently desegregated South[96] and the denunciation of the examinations by the National Education Association focused attention on test validity and use. The yearly volume of candidates had more than doubled since the beginning of the decade, growing from the 37,000 tested in 1959-60[97] to almost 73,000 in 1963-64.[98] Much of this growth occurred in the South.[99] In 1963-64, eighty-one percent of those registering to take the exam at a nationwide administration resided in the South Atlantic or South Central regions of the

country.[100] By 1968, the exams were required of all candidates in South Carolina, North Carolina, Texas, and West Virginia and of applicants of the "grants-in-aid" program in Georgia. Additionally, they were often required locally in the District of Columbia, Maryland, Virginia, Georgia, Arkansas, Louisiana, and Oklahoma.[101] For many southern community and state school systems, the decade of the 1960's was a period of considerable turmoil and change, much of which was in response to court-ordered desegregation.[102] In a number of these school systems, a related change was an increased reliance upon the National Teacher Examinations.

Over time, test use became subject to greater and greater critical attention both within and outside of the Educational Testing Service. In 1966, the National Education Association resolved "that the use of examinations such as the National Teachers Examination [was] not a desirable method of evaluating teachers in service . . ."[103] By 1970, it had strengthened its position against the exams and resolved "that examinations such as the National Teacher Examinations must not be used as a condition of employment or a method of evaluating educators in service for purposes such as salary, tenure, retention, or promotion."[104] In the early 1970's, the National Education Association joined the U.S. Justice Department in several court challenges[105] in which "black educators in the deep South contended that [the exam's] use had a racially discriminatory effect on minority employment in the public schools."[106]

The Educational Testing Service and its advisory groups on teacher examinations responded to the concerns and criticisms in several ways. Formal guidelines for proper use were developed throughout the 1960's and were distributed to test users in 1971. Existing tests were carefully scrutinized --both experimentally[107] and with the review and revision of the test specifications. In 1969, a panel of minority group educators was invited to review the exams and make suggestions. The following year the tests were modified in response to the panel's suggestions, most of which dealt with the content of specific sections. Beginning with the 1970-71 administrations, the test structure was as follows:

General Education	61 percent
1. Written English Expression	11 %
2. Social Studies, Literature, and the Fine Arts	25%
3. Science and Mathematics	25%
4. Professional Education	39 percent

The Impact of Judicial Interpretations

About this time, legal challenges to the use of other employment and licensing tests focused new attention to issues of validity and validation. In March 1971, in Griggs v. Duke Power Company,[108] the Supreme Court reinforced policies established by the U.S. Equal Opportunity Commission the previous year. In the first of several landmark cases, the high court ruled that employment tests, with a disproportionate exclusionary impact on groups protected by the Civil Rights Act of 1964, must be "shown to be related to job performance."

Several months after the decision, James Deneen, then ETS's director of teacher examinations, issued a statement on the ruling's impact on teacher testing. He argued that the National Teacher Examinations were "job-related in so far as they measure knowledge that is needed and applied in teaching. The tests' specifications and questions are prepared by specialists who teach the subjects examined at the college and university level and by school district teachers, supervisors, and administrators. The factors and items found in the NTE are based on teacher training programs. Thus the tests possess content validity, which is basic to any achievement test."[109]

Deneen wrote of the content review by black educators and of the plans to add "more items which reflect the contributions of minority groups." In argumentation very similar to that recently used by ETS president Gregory Anrig[110] and others who defend the use of the teacher tests despite their documented negative impact on minority teachers, Deneen wrote: "Most black teacher trainees who take the NTE are products of segregated colleges, segregated elementary and high schools, and segregated neighborhoods. They are largely drawn from a population which has possessed little economic, social, or political power to change its educational environment. It is obvious that, regardless of their race, persons with such a background will generally score lower on an educational achievement test than their more privileged colleagues. It seems equally obvious that the appropriate response to this fact is not to depreciate the importance of knowledge for teachers, but to make that knowledge available to all regardless of race or socioeconomic status."[111] Stating that the "Court's decision [pointed] up the urgency of developing more and better criteria for measuring teaching," Deneen also described some of the validation work then underway at ETS.

Over the next few years, considerable internal attention was paid to issues of validity and validation. Previous research was re-evaluated[112] and new procedures were considered. Guidelines were provided so that users could conduct their own

validation studies,[113] and self-reported grade point data were gathered during testings to further explore concurrent validity.[114] The most ambitious and most decisive validation study conducted by ETS in the 1970's was that undertaken for the State of South Carolina. It was this study which slowed the tide of court cases filed against the use of the National Teacher Examinations and established current NTE validation procedures.

In 1975, the United States Department of Justice, the National Education Association, and groups of South Carolina's teachers charged that the use of the National Teacher Examinations in South Carolina for teacher certification and as a factor in determining salary violated the equal protection clause of the Fourteenth Amendment and Title VII of the Civil Rights Act of 1964. In January 1978, the United States Supreme Court refused to accept the case for full briefing and oral argument[115] and summarily affirmed the 1977 decision of the Federal District Court[116] which stated: "The State has the right to adopt academic requirements and to use written tests designed and validated to disclose the minimum amount of knowledge necessary to effective teaching."[117] "There is ample evidence in the record of the content validity of the NTE. The NTE have been demonstrated to provide a useful measure of the extent to which prospective teachers have mastered the content of their teacher training programs."[118]

A good deal of the courts' faith in the content validity of the tests was based on the study conducted by the Educational Testing Service for the South Carolina Department of Education.[119] About 450 faculty members from some twenty-five teacher training institutions in South Carolina examined the test items to determine if they fairly sampled the knowledge which the teacher training institutions sought to impart. Content review panels judged "whether or not the content of each question . . . [was] covered by the teacher education program" and assessed "the relation between the description of test content . . and the curriculum in terms of omission or overemphasis."[120] Knowledge estimation panels provided "estimates of the percentages of minimally knowledgeable candidates who would be expected to know the answers to individual test questions."[121] Thus, faculty members' judgments as to the minimum amount of knowledge needed to complete a South Carolina teacher education program were used to calculate cutoff scores for the common exam and each of the area exams.

In the next few years, the NTE were validated--using the South Carolina model--for certification in California, Louisiana, and North Carolina and for licensure by the American Speech-Language-Hearing Association.[122] In each case, test items were compared with curricula of the teacher training institutions and

the conclusion reached that the tests were "valid" because they were representative of the content taught in those programs.

The Policy Council and Its New Core Battery: The NTE's in the 1980's

In the current decade, the Educational Testing Service and those responsible for the NTE program have tried to avoid the unflattering controversy and costly legal entanglement of the past while profiting from a market created by an intense public demand for teacher testing. In 1979, ETS selected a twelve person external board, the "National Teacher Examinations Policy Council," to govern and direct NTE program policies involving the development, administration, and use of the exams.[123] Members, who beginning in 1982 also included classroom teachers, were drawn from states and school districts that used the tests and from user and non-user institutions of higher education and were appointed in order to "make the program more responsive to user requirements."[124] Created to insulate ETS from controversial policy and legal decisions, the Policy Council was given "all policy making responsibility" for the NTE program.[125] ETS personnel, however, continue to take responsibility--and credit--for popular actions such as the decision to disallow NTE sales for the testing of experienced teachers in Arkansas.[126]

Reiterating that "the basic purpose of the tests" was "to provide a measure of academic preparation for beginning teachers,"[127] the Policy Council introduced a major NTE revision in the fall of 1982. Criticisms of the previous tests and the "screening, counseling, guidance, and feedback needs of teacher education institutions" were taken into consideration. Consisting of three distinct sections to be administered at the same time or in separate two-hour blocks, the new Core Battery samples content similar to that covered in the earliest National Teacher Examinations--basic communicative skills, cultural background, and professional information. This time, however, the three tests are scored and reported separately and are not combined into a single score.[128] Test names, sections, and components[129] are as follows:

Test of Communication Skills

1. Listening ..... 40 multiple choice questions
2. Reading ..... 30 multiple choice questions
3. Writing ..... 45 multiple choice questions
4. Writing ..... one essay question

Test of General Knowledge

1. Social Studies .....30 multiple choice questions
2. Mathematics ..... 25 multiple choice questions
3. Literature and Fine Arts .....35 multiple choice questions
4. Science ..... 30 multiple choice questions

Test of Professional Knowledge

4 sections of 35 multiple choice questions each, only 3 of which are scored.

Careful to operate within the bounds of past court decisions, ETS personnel initially stressed that because "the Core Battery was sufficiently different from the Common Examinations, the qualifying scores established for the Commons [could not] be used with the Core Battery Tests."[130] Thus score users were advised that it would be necessary that they conduct new validity studies "to examine the relationship of the new test content to what is taught . . ."[131]

Validation of "what is taught" is, however, no longer legally sufficient. Responding to previous legal challenges and to the 1978 adoption of the Equal Employment Opportunity Commission's Uniform Guidelines on Employee Selection Procedures,[132] the latest NTE use guidelines require that the "NTE Program tests be validated for the specific purposes for which they are being used."[133] They point out that "in addition, federal and other civil rights laws, such as Title VI and Title VII of the Civil Rights Act of 1964, may also require validation if the use being made of the tests is shown to disproportionately disadvantage members of ethnic, racial, religious, or gender subgroups."[134] Users are referred to appropriate "professional and legal standards" and are advised that "in some cases these standards require the use of job analyses or other similar techniques."[135]

Earlier this year, ETS president Gregory Anrig announced that a recent "job analysis project" which involved the participation of some 3000 classroom teachers will soon be published and will be used in the future "to assist in developing and validating the NTE for state certification."[136] To date, however, validation of the Core Battery has involved judging procedures similar to those followed in the South Carolina study. In addition to considering the tests' similarity to teacher education curricula, judges--who may be teacher education personnel,[137] or practicing teachers,[138] or both[139]--are now asked to compare test items to that knowledge required by beginning or minimally qualified teachers.

Conclusion

Three major findings emerge from historical consideration of the validity and validation of the National Teacher Examinations. These relate to (1) the continuity of test content and justification over the fifty year period of the program's existence, (2) the primacy of reliance upon logical or content validity, and (3) the paradoxical relationship of the tests to teacher education curricula.

First, since their inception the National Teacher Examinations have measured three categories of teacher knowledge--basic intellectual and communicative skills, general cultural and contemporary background, and pedagogical and professional information. Although the exam evolved from the comprehensive multi-sectioned battery administered in 1940 to the more narrowly focused tests of the 1960's and 1970's and then to the three part core of the 1980's, clearly, that first test set the pattern for those which followed it. There has been a strong tendency to maintain the status quo and to continue relying upon previous models of the exams even when those models were "inherited" from prior agencies or test developers. No additional or innovative sections were adopted until 1982.

When changes did occur, they were undertaken for either financial reasons or as responses to specific criticisms. Changes made in the 1940's and 1950's were undertaken to save construction and administration costs and those made in the 1960's reflected the criticisms of the NCTEPS and minority group review committees. Certainly, the restructuring of the exam in 1982 responded both to previous criticisms and to a perceived new test market. Throughout this evolution, however, the official justification for test use has continued to focus upon the perceived incompetence of many teachers and the assumed inadequacy of their training.

Second, there has been a strong and persistent tendency to justify the exams in terms of their logical or practical validity. Statistical validation was de-emphasized and even ridiculed by many of those in charge of the program until necessitated by the courts. Again and again, test validation was justified using arguments and strategies which were no longer appropriate or true. Reliance upon explanations of the tests' careful construction persisted through periods in which actual construction was inadequate. Justification based upon the inadequacy of existing teacher training was used even when those being tested had long been away from the training institutions.

And finally, despite the persistent assumption that the tests are needed because graduates of many teacher education programs are inadequately prepared, the source of test content and validity has been and continues to be focused primarily upon the perceived curricula of those programs.

The original tests assessed knowledge selected by administrators as representative of what all "good teachers should know." Program officials stressed that the exams did not measure the totality of teaching ability and therefore should not be judged by their correlation with "available criteria of teaching success." Despite this, a number of early studies attempted to demonstrate the exam's predictive validity by comparing test scores with supervisors' or principals' ratings. Later, investigations were broadened to include comparisons with concurrent measures of achievement in college or graduate school. Program personnel have tended to attribute the low to moderate correlations yielded by these studies to their theoretical or technical inadequacies.

By the mid 1960's, the examinations were said to appraise "basic professional preparation and general academic attainment." Content validity--justified primarily in terms of the qualifications of those nationally selected and recognized experts who assisted in test development--was emphasized in the program's publications.

Beginning in the early 1970's, a number of law suits charged that the tests were being used in some states and communities to discriminate against minority teachers and teacher candidates. Teachers' unions and other critics claimed that the tests were inappropriate because they were not "validated" against job-related criteria. In 1978, the U.S. Supreme Court ruled in favor of the exams' use and thus indirectly in favor of their content validity. Although practicing teachers are now asked to judge the relevance of the items, current validation procedures tend to closely mirror those used in the South Carolina study and focus upon the similarity of test content to the curricula of training institutions. This is the case despite the prevalent assumption that testing is now justified, as it was fifty years ago, on the basis of these institutions graduating inadequately trained teachers.

Notes

[1]Ann Jarvelle Wilson, "Knowledge for Teachers: The National Teacher Examinations Program, 1940-1970," unpublished Ph.D. dissertation, University of Wisconsin, 1984, available from University Microfilm International as No. 84-14265.

[2]Access to the NTE history files was provided by the Educational Testing Service, and permission to cite excerpts from those files was granted by ETS and the American Council on Education--sponsor of the NTE program from 1940 to 1949. The opinions and conclusions reached in this paper are those of the author, and neither ETS nor ACE has participated in or bears any responsibility for this study.

[3]Robert L. Ebel, "The Practical Validation of Tests of Ability," Educational Measurement: Issues and Practice 2 (Summer 1983), p. 7.

[4]C.W. Odell, Traditional Examinations and New-Type Tests (New York: Century Co., 1928), p. 59.

[5]John C. Flanagan, "The Use of Comprehensive Rationales in Test Development," Educational and Psychological Measurement 11 (Spring 1951), pp. 151-155.

[6]Walter Monroe, "Some Trends in Educational Measurement," in Twenty-Fourth Annual Conference on Educational Measurement (Bloomington: Indiana University Press, 1937), p. 33. Also see discussion in Walter Haney, "Validity, Vaudeville, and Values" A Short History of Social Concerns over Standardized Testing," American Psychologist 36 (October 1981), pp. 1021-1034.

[7]Lyle F. Schoenfeldt, "The Status of Test Validation Research," in Social and Technical Issues in Testing: Implications for Test Construction and Usage, Barbara Plate, ed. (Hillsdale, NJ: Lawrence Erlbaum, 1984), p. 64.

[8]American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing (Washington, DC: American Psychological Association, 1985), p. 10.

[9]Often these were those also involved in test construction. For example, see Ebel, 1983 and John C. Flanagan, "A Rational Rationale," Educational Measurement: Issues and Practice 2 (Summer 1983), p. 12.

[10]For example, see discussion in Nancy Cole, "Bias in Testing," American Psychologist 36 (October 1981), pp. 1067-1077.

[11]American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education, Technical Recommendations for Achievement Tests, (Washington, DC: American Psychological Association, 1955.

[12]Ibid., p. 16.

[13]American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, Standards for Educational and Psychological Tests and Manuals, (Washington, DC: American Psychological Association, 1966.

[14]Schoenfeldt, p. 65.

[15]See, for example, Samuel Messick, "The Standard Problem: Meaning and Values in Measurement and Evaluation," American Psychologist 30 (October 1975), pp. 955-966 and Melvin R. Novick, "Importance of Professional Standards in Fair and Appropriate Test Use," in The Uses and Misuses of Tests, Charles W. Deves, ed., (San Francisco: Jossey-Bass, 1984), pp. 13-19.

[16]AERA et al., Standards, 1985, p. 10.

[17]Lee Cronbach, "Validity on Parole: How Can We Go Straight?," in New Directions for Testing and Measurement: Measuring Achievement Progress Over a Decade, W. Schroder, ed. (San Francisco: Jossey-Bass, 1980), p. 99.

[18]See discussion in Walt Haney, "Testing Reasoning and Reasoning about Testing," Review of Educational Research 54 (Winter 1984), pp. 597-654 and in David Resnick, "Testing in America: A Supportive Climate," Phi Delta Kappan 62 (May 1981), pp. 625-628.

[19]Wilson, "Knowledge for Teachers," pp. 42-57. Also see David Resnick, "History of Educational Testing," in Ability Testing: Uses, Consequences, and Controversies, Part II, A.K. Wigdor and W. R. Gardner, eds. (Washington, DC: National Academy Press, 1982), pp. 173-174.

[20]For example, one large scale teacher assessment project was promoted in the late 1920's by the Bureau of Public Personnel Association and another in the early 1930's by the Teachers College Personnel Association. [Wilson, "Knowledge for Teachers," pp. 57-59.]

[21]William Learned and Ben Wood, The Student and His Knowledge: A Report to the Carnegie Foundation on the Results of the High School and College Examinations of 1928, 1930, and 1932, Carnegie Foundation for the Advancement of Teaching Bulletin No. 29 (New York: D.B. Updike, Merrymount Press, 1938.)

[22]Ibid., pp. 15-17.

[23]An average correlation of .63 was obtained at the sixteen colleges studied in 1928, and those of .89-.97 were calculated at one college "where it [was] known that the professors [were] well acquainted with the students and [exercised] great care in assigning grades." [Ibid.]

[24]Wood, Ben D., "Teacher Selection: Tested Intelligence and Achievement of Teachers-in-Training," Educational Record 17 (July 1936), pp. 374-387.

[25] Learned and Wood, p. 43.

[26]Ibid., p. 65.

[27]Ben D. Wood, "Ten Years of the Cooperative Test Service," Educational Record 21 (July 1940), pp. 368-380.

[28]Howard Long et al., Principles and Procedures of Teacher Selection (Philadelphia: American Association of Examiners and Administrators of Educational Personnel, 1951), p. 15.

[29]Donald J. Shank, "Minutes of the Meeting of the National Committee of Teacher Examinations, September 23-24, 1939, New York City," confidential unpublished document from NTE History File.

[30]John Flanagan, "An Analysis of the Results from the First Annual Edition of the National Teacher Examination," Journal of Experimental Education 9 (March 1941), p. 258.

[31]From the beginning, the National Teacher Exams have consisted of two separate tests--a "common" exam of general and professional knowledge for all teachers and separate subject-matter tests for specific teacher groups.

[32]Shank, p. 5.

[33]Ben D. Wood, An Announcement of a Teacher Examination Service (New York: National Committee on Teacher Examinations of the American Council on Education, 1939), p. 4.

[34]A.J. Stoddard, "The Selection of Teachers from the National Viewpoint," Educational Record 21 (January 1940), p. 151.

[35]Wood, "An Announcement," p. 12.

[36]Ben D. Wood, "Making Use of the Objective Examination as a Phase of Teacher Selection," Harvard Educational Review 10 (May 1940), p. 277.

[37]Wood, "Making Use," pp. 277-279.

[38]See for example, Albert Lindsey Rowland, "The Proposed Teacher Education Service," Harvard Educational Review 10 (May 1940), pp. 283-288 and Harold Hand, "Hazards in National Examinations," Frontiers of Democracy 6 (May 15, 1940), pp. 228-229.

[39]W. Carson Ryan, "National Examinations for Teachers?," Progressive Education 16 (December 1939), p. 531.

[40]John Pilley, "The National Teacher Examination Service," School Review 49 (March 1941), p. 179.

[41]David G. Ryans, "The National Teacher Examinations: Notes on the Question of Their Validity," (New York: National Committee on Teacher Examinations of the American Council on Education, 1946), p. 1.

[42]Ben D. Wood, "Dr. Wood's Statement," Progressive Education 17 (March 1940) p. 156.

[43]Wood, "Making Use," p. 279.

[44]David G. Ryans, "The Professional Examination of Teaching Candidates: A Report of the First Annual Administration of the National Teacher Examinations," School and Society 52 (October 5, 1940), p. 275.

[45]John C. Flanagan, "A Preliminary Study of the Validity of the 1940 Edition of the National Teacher Examinations," School and Society 54 (July 1941), pp. 59-64.

[46]Flanagan, "The Use of Comprehensive Rationales," and Ebel, "The Practical Validation."

[47]Flanagan, "A Preliminary Study," pp. 59-60.

[48]Ibid., pp. 62-63.

[49]For example, see Leo Joseph Lins, "The Prediction of Teaching Efficiency," Journal of Experimental Education 15 (September 1946), pp. 2-60; and David G. Ryans, "The Results of Internal Consistency and External Validation Procedures Applied in the Analysis of Test Items Measuring Professional Information," Educational and Psychological Measurement 11 (Winter 1951), pp. 549-560.

[50]For example, see John T. Dailey, "Development and Applications of Tests of Educational Achievement Outside the Schools," Review of Educational Research 23 (February 1953), pp. 102-109 and Thomas J. Quirk et al., "Review of Studies of the Concurrent and Predictive Validity of the National Teacher Examinations," Review of Educational Research 43 (1973), pp. 89-113. Also see Ralph W. Tyler, "The Specific Techniques of Investigation: Examining and Testing Acquired Knowledge, Skill and Ability," in National Society of the Study of Education, Thirty-Seventh Yearbook (Bloomington, Indiana: Public School, 1938), pp. 341-355.

[51]See, for example, "Study of N.T.E. Results at 'X' College," in The Selection of Teachers, Bulletin No. 1 (New York: National Committee on Teacher Examinations of the American Council on Education, August 27, 1946), p. 3 and May V. Seagoe, " The Prediction of Success in a Graduate School of Education," School and Society 69 (February 5, 1949), pp. 89-93.

[52]Both Ben Wood and John Flanagan left their Cooperative Test Service positions in the early 1940's for war related work. Ryans, who had been the program's executive secretary earlier, became director of the testing program of the Cooperative Test Service.

[53]David G. Ryans, "National Committee on Teacher Examinations: Considerations for the Future," October 28, 1944, confidential unpublished document from the NTE History File, p. 2.

[54]"National Committee On Teacher Examinations, "Minutes" of the November 1, 1944 meeting in Philadelphia, confidential unpublished document from the NTE History File, p. 1.

[55]Wilson, "Knowledge for Teachers," pp. 197-200.

[56]David G. Ryans, "Summary Report: 1946 National Teacher Examinations Program," Teacher Selection Papers and Reports, No. 9 (New York: National Committee on Teacher Examinations of the American Council on Education, January 17, 1947), pp. 5-6.

[57]National Committee on Teacher Examinations, "Minutes" of the November 1, 1944 meeting in Philadelphia, confidential unpublished document from the NTE History File, pp. 2-3.

[58]For example, David G. Ryans, "Use of the National Teacher Examinations in Colleges and Universities," Teacher Selection Papers and Reports, No. 6 (New York: National Committee on Teacher Examinations of the American Council on Education, October 9, 1946).

[59]Howard J. Savage, "Educational Grants of the Carnegie Corporation and the Carnegie Foundation: Projects Terminated," in Carnegie Foundation for the Advancement of Teaching, Fortieth Annual Report, 1944-45 (New York: D.B. Updike, Merrymount Press, 1945), p. 41-42.

[60]Ann Jervelle Wilson, "Historical Questions of Equity and Excellence: South Caroline's Adoption of the National Teacher Examinations," manuscript submitted in November 1985 for publication to Urban Educator.

[61]J. McTyiere Daniel, Excellent Teachers: Their Qualities and Qualifications (Columbia: Steering Committee of the Investigation of Educational Qualifications of Teachers in South Carolina, University of South Carolina, 1944).

[62]Ibid., p. 238.

[63]Wilson, "Historical Questions of Equity."

[64]E.R. Crow, "Teacher Examinations and the South Caroline Certification Program," Teacher Selection Papers and Reports, No. 8 (New York: National Committee on Teacher Examinations of the American Council on Education, January 1, 1947. An edited version with the same title but with racial inferences removed appeared in Educational Record 28 (October 1947), pp. 454-462.

[65]David G. Ryans, "The Function of Examinations in the Selection of Teachers," School Executive 68 (May 1949), pp. 39-41.

[66]"Study of N.T.E. Results at 'X' College."

[67]Ryans, "The NTE: Notes on the Question of Their Validity," p. 3.

[68]Ibid., p. 4.

[69]"ETS and the Test of Time: A 25-Year Review," in Educational Testing Service, Annual Report, 1973, 1974: Flexibility for the Future (Princeton: ETS, 1974), p. 11.

[70]Henry Chauncey, "National Teacher Examinations Program," in Educational Testing Service, Annual Report to the Board of Trustees, 1949-50 (Princeton: ETS, 1950, p. 71.)

[71]"Teacher Examination Activities, 1940-57," (Princeton: ETS, January 19, 1959), confidential unpublished document from the NTE History File, p. 2.

[72]Wilson, "Knowledge for Teachers, " p. 239.

[73]See, for example, H. Natalie Sutcliffe, "A Study of Some Correlations Existing between the Four Year Indices of Selected Seniors and Graduates-in-Service of the Rhode Island College of Education and Their National Teacher Examinations Scores: 1940-1950," unpublished Master of Education thesis, Rhode Island College of Education, June 1953 and James E. McCamey, Jr., "The Correlations between Certain Academic Factors and Scores of the 1957 National Teacher Examinations of the 1957 Graduates of the University of Hawaii Teachers College," unpublished Master of Education thesis, University of Hawaii, June 1958.

[74]Joseph Augustine Shea, "The Predictive Value of Various Combinations of Standardized Tests and Subtests for Prognosis of Teaching Efficiency," in Catholic University of America, Educational Research Monographs, Volume 19, Number 5 (Washington, D.C.: Catholic University of America Press, June 1955).

[75]See Eleanor Cecilia DeLaney, "Teacher Selection and Evaluation: with Special Attention to the Validity of the Personal Interview and the National Teacher Examinations as Used in One Selected Community (Elizabeth, New Jersey)," [unpublished Ph.D. dissertation, Columbia University, 1954], abstract in Dissertation Abstracts, Volume 14 (Ann Arbor, Michigan: University Microfilms, 1954), pp.1334-1335 or Helen Murray Kleyle, "Differences in Personal and Professional Characteristics of a Selected Group of Elementary Teachers with Contrasting Success Records," [unpublished Ph.D. dissertation, University of Pittsburgh, 1959], abstract in Dissertation Abstracts, Volume 20 (Ann Arbor: University Microfilms, 1959), pp. 185-186.

[76]Educational Testing Service, The National Teacher Examinations: Handbook for School and College Officials (Princeton: ETS, 1959), p. 13.

[77]Ibid., p. 11.

[78]Ibid., page 15.

[79]Ibid.

[80]For example, see Arthur L. Benson, "Testing for Professional Selection," in National Council on Measurements Used in Education, Sixteenth Yearbook (East Lansing: the Council, 1959), pp. 26-30.

[81]Arthur L. Benson, "The Role of Examinations in the Preparation of Teachers," Journal of Teacher Education 10 (December 1959), p. 492.

[82]Arthur L. Benson, "Testing Procedures in the Administration of Educational Personnel," Education 75 (December 1954), p. 244.

[83]Educational Testing Service, The National Teacher Examinations: Handbook for School and College Officials (Princeton: ETS, 1961), p. 15.

[84]Educational Testing Service, The National Teacher Examinations: Prospectus for School and College Officials (Princeton: ETS, 1964), p. 17.

[85]Educational Testing Service, Prospectus for School and College Officials: The National Teacher Examinations, Princeton: ETS, 1967, p. 19.

[86]For example, in his book, The Miseducation of Teachers, James Koerner called for teacher testing but decried the quality of the National Teacher Examinations since they assessed "only what educationists tell ETS they are trying to do in their teacher training programs." [James Koerner, The Miseducation of Teachers (Baltimore: Penguin Books, 1963), pp. 254-255.

[87]"Recommended Revisions of the National Teacher Examinations: A Report of Suggestions Made by the NTE Review Committee," (Princeton: ETS, March 13, 1961), p. 6.

[88]Ibid., pp. 2-3.

[89]ETS, The NTE: Prospectus, 1964, p. 5.

[90]Ibid., p. 17.

[91]Educational Testing Service, Technical Handbook: The National Teacher Examinations (Princeton: ETS, 1965), p. 18.

[92]Barbara Pitcher, "The Relationship of Academic Success in Teacher Preparatory Curricula to Scores on the NTE Common Examinations," ETS Statistical Report SR-62-63 (Princeton: ETS, November 1962.)

[93]Betty Humphry, "A Survey of Professional Education Offerings in NCATE-Accredited Institutions," Journal of Teacher Education 14 (1963), pp. 406-410.

[94]For example, see "The National Teacher Examinations: Notes on the Question of Their 'Validity'," (Princeton: ETS, April 1978), p. 3 and Educational Testing Service, National Teacher Examinations: Technical Handbook (Princeton: ETS, 1973), p. 12.

[95]Humphry, p. 409.

[96]Increased investigation via doctoral dissertations accompanied this use. Especially see Ernest C. Phillips, Jr., "A Comparative Study of the Performance of White and Negro Teachers on the Individual Items of a Standardized Test of Teaching Competence," unpublished Ed.D. dissertation, University of Georgia, 1956; Hazel Deal Simpson, "An Analysis of the Relationship between Scores Attained on the National Teacher Examinations and Certain Other Factors," unpublished Ed.D. dissertation, University of Georgia, 1962; and S. J. Tullos, "An Investigation of the Uses of the National Teacher Examinations," unpublished Ed.D. dissertation, Colorado State College, 1967. Also, see, James E. Greene, "A Comparison of Certain Characteristics of White and Negro Teachers in a Large Southeastern School System," Journal of Social Psychology 58 (December 1962), pp. 383-391.

[97]"Descriptive Listing of Testing Programs," in Educational Testing Service, Annual Report, 1959-60 (Princeton: ETS, 1960), pp. 56-57.

[98]"Descriptive Listing of Testing Programs," in Educational Testing Service, Annual Report, 1963-64 (Princeton: ETS, 1964), pp. 80-82.

[99]Although apparently not by other ETS officials, Southern use certainly had been encouraged by Arthur Benson. In 1954, shortly after the Brown decision by the U.S. Supreme Court, he had pointed out to Southern school officials that black and white teachers tended to score differently on the teacher examinations and suggested that with the use of the exams "the South [could] face its future with confidence . . ." [Arthur L. Benson, "Problems of Evaluating Test Scores of White and Negro Teachers," in Proceedings of the Southern Association of Colleges and Secondary Schools (Atlanta: the Association, 1955), p. 176.

[100]"Volume of NTE Candidates by State for February 16, 1963 Nationwide Administration," confidential unpublished document from NTE History File. These regions, as defined in the Office of Education's "biennial surveys of education," included the states of Delaware, Maryland, Virginia, West Virginia, North Carolina,South Caroline, Georgia, Floride, Kentucky, Tennessee, Alabama, Mississippi, Arkansas, Louisiana, Oklahoma, and Texas as well as the District of Columbia.

[101]Arthur L. Benson, "Summary of Recent Activities in the National Teacher Examinations" (Princeton: ETS, May 1968), confidential unpublished document from the NTE History File, p.#2.

[102]See Richard Kluger, Simple Justice: The History of Brown v. the Board of Education and Black America's Struggle for Equality (New York: Vintage Books, 1975.)

[103]"Resolution 66-11 on Evaluation and Subjective Ratings," in National Education Association, Addresses and Proceedings, 1966, Volume 104 (Washington, D.C.: NEA, 1966), p. 471.

[104]"Resolution C-6 on Evaluation and Subjective Ratings, " in National Education Association, Addresses and Proceedings, 1972, Volume 110 (Washington, D.C. : the NEA, 1972), pp. 678-679.

[105]These first cases were in Mississippi and Alabama. [Baker v. Columbus Municipal Separate School District, 329 F. Supp. 706 (Mississippi, 1971), affirmed in 462 F. 2nd 1112 (5th Cir. 1972) and Lee v. Macon County Board of Education, 463 F. 2nd 1174 (5th Cir. 1972).

[106]National Education Association, Teacher Rights Division, "Testing Teachers: An Unfair Game," Today's Education 61 (October 1972), p. 60. Also see Betty E. Sinowitz, "The Teacher and the Law: Teachers Fight Misuse of National Teacher Exams," Today's Education 65 (January-February 1976), pp. 108-109.

[107]See, for example, Thomas J. Quirk and Donald M. Medley, "Race and Subject-Matter Influences on Performance on General Education Items of the National Teacher Examinations," Proceedings of the Eightieth Annual Convention of the American Psychological Association 7 (1972), pp. 469-470.

[108]Griggs v. Duke Power Company, 401 U.S. 424 (1971).

[109]James R. Deneen, "Tests and Teacher Performance: The Impact of the Griggs Opinion," (Princeton: ETS, September 1971), p. 1.

[110]Gregory R. Anrig, "Teacher Education and Teacher Testing: The Rush to Mandate," Phi Delta Kappan 67 (February 1986), pp. 447-451.

[111]Deneen, p. 3.

[112]For example, see Thomas J. Quirk et al., "The National Teacher Examinations: An Annotated Bibliography, 1940-1971," ETS Research Memorandum RM-72-4 (Princeton: ETS, April 1972) and Thomas J. Quirk et al., "Review of Studies of the Concurrent and Predictive Validity of the National Teacher Examinations," Review of Educational Research 43 (1973), pp. 89-113.

[113]Thomas J. Quirk, "A Manual for Designing and Conducting Validity Studies Based on the National Teacher Examinations," ETS Report PR-72-8 (Princeton: ETS, May 1972), available as ERIC Document ED 068 577.

[114]Norman Wexler, "Concurrent Validity of the National Teacher Examinations," March 1975, available as ERIC Document ED 110 477.

[115]National Education Association v. South Caroline and United States v. South Caroline, 434 U.S. 1026, 98 S. Ct. 756 (1978).

[116]United States v. South Caroline, 445 F. Supp. 1094 (S.C. 1977).

[117]Cited in ETS Information Division, "U.S. Supreme Court Acts on N.E.A. v. South Caroline: Use of the N.T.E. Upheld, in NTE News (Princeton: ETS, January 1978), p. 4.

[118]Ibid., p. 3.

[119]Educational Testing Service, "Report on a Study of the Use of the National Teacher Examinations by the State of South Caroline, prepared for submission to South Caroline Department of Education, Columbia, South Caroline" (Princeton: ETS, January 1, 1976). Also see discussion in Thomas R. McDaniel, "The NTE and Teacher Certification," Phi Delta Kappan 59 (November 1977), pp. 186-188.

[120]Memorandum to panel members from the state superintendent, Fall 1975. Cited in McDaniel, p. 188.

[121]Ibid.

[122]ETS Information Division, "Standard Study Design Refined to Examine NTE Validity," in NTE News (Princeton: ETS, January 1979), p. 4.

[123]ETS Information Division, "NTE Policy Rests with New Council," in NTE News (Princeton: ETS, January 1979), p. 1.

[124]Frieda C. Rosner and David R. Krathwohl, "Teacher Certification and the NTE," Educational Measurement: Issues and Practice 1 (Summer 1982), p. 24.

[125]Ibid.

[126]See Jim Bencivenga, "The Flap over Teacher Tests Takes a New Twist," Christian Science Monitor, December 16, 1983, p. 23 and Barbara Burgower, "Taking the Test for Teachers," Newsweek, January 9, 1984, p. 82. Interestingly, it was later stated that both "ETS and the NTE Policy Council" had disallowed this use. [Anrig, p. 449.]

[127]Rosner and Krathwohl, p. 24.

[128]ETS, NTE Programs: Core Battery Tests (Princeton: ETS, 1982).

[129]NTE Policy Council, A Guide to the NTE Core Battery Tests: Communication Skills, General Knowledge, Professional Knowledge (Princeton: ETS, 1984), pp. 133-138.

[130]William Harris, NTE Newsletter (Princeton: ETS, August 1984), p. 2.

[131]Ibid.

[132]These specifically call for validation in terms of "job performance." See discussion in Rodney Roth, "Validation Study of the National Teacher Examinations for Certification in the State of Arkansas," paper presented as part of a symposium on "Validation of the National Teacher Examinations: A Multi-State Perspective" at the Annual Meeting of the American Educational Research Association in New Orleans, April 1984.

[133]NTE Policy Council, NTE Program Guidelines for Proper Use of NTE Tests (Princeton: ETS, 1985), p. 9.

[134]Ibid.

[135]Ibid.

[136]Anrig, p. 449.

[137]For example, see Brenda J. Hankins and James J. Hancock, "Validation of the NTE for Certification of Entry Level Teachers in the State of Mississippi," paper presented as part of a symposium on "Validation of the National Teacher Examinations: A Multi-State Perspective" at the Annual Meeting of the American Educational Research Association in New Orleans, April 1984.

[138]For example, see Carroll Hall, "Validating the NTE for the Initial Certification of Teachers and Administrators in New Mexico . . . and Beyond," paper presented as part of a symposium on "Validation of the National Teacher Examinations: A Multi-State Perspective" at the Annual Meeting of the American Educational Research Association in New Orleans, April 1984.

[139]Roth, "Validation Study of the National Teacher Examinations for Certification in the State of Arkansas."